机器人遵从伦理促进人机信任?
决策类型反转效应与人机投射假说

王 晨[1] 陈为聪[2,3] 黄 亮[2,3] 侯苏豫[2,3] 王益文[1]

([1] 福州大学经济与管理学院, 福州 350116)([2] 闽南师范大学应用心理研究所, 漳州 363000)

([3] 福建省应用认知与人格重点实验室, 漳州 363000)

摘 要 阿西莫夫三大伦理原则是关于人工智能机器人的基本伦理规范。本研究提出人机投射假说——人会从自身具有的认知、情感和行动智能出发, 去理解机器人的智能并与之互动。通过三个实验, 从原则一到原则三逐步考察在机器人是否遵守伦理原则对人机信任的影响中, 机器人决策类型(作为与否; 服从人类命令与否; 保护自身与否)的效应, 以及人机投射的潜在机制。结果揭示了人机投射在机器人遵守伦理原则促进人机信任中起中介作用, 以及机器人决策类型与是否遵守伦理原则之间有趣且有意义的交互效应: (1)在遵守情境下, 机器人作为相对于不作为更有利于促进信任, 但在违反情境下, 则反之; (2)在遵守且尤其在违反情境下, 机器人服从相比不服从人类命令更有利于促进人机信任; (3)相较于违反情境, 机器人保护相比不保护自身在遵守情境下更有利于促进人机信任。跨实验的分析更深入地阐释了在遵守和违反伦理原则情境中以及伦理要求冲突情境中, 有利于促进人机信任的机器人行动决策因素。

关键词 人工智能, 机器人伦理原则, 人机信任, 人机投射, 人机交互

1 前言

唯一的保障是制造出来的机器人总是严格地遵守第一原则——机器人在任何情况下都不能伤害人类。

——Asimov (1942)

随着人工智能(artificial intelligence, AI)技术的蓬勃发展, 机器人作为人工智能的典型实例逐渐被应用到了生产、医疗、教育乃至军事等领域中, 在促进人类福祉的同时也具有潜在的伦理风险。例如, 机器人可能在人机交互中对人类造成身体或心理伤害(Maninger & Shank, 2022), 机器算法决策可能导致种族、年龄和性别歧视问题(Bigman & Gray, 2018), 抑或是机

收稿日期: 2022-07-04

* 国家社会科学基金重大项目(19ZDA361)和国家社会科学基金青年项目(20CSH069)阶段性成果。

通信作者: 王益文, E-mail: wangeven@126.com; 黄亮, E-mail: yeoo5860@163.com

器生产、自动驾驶技术或将使数百万人失去工作(Etemad-Sajadi et al., 2022)，甚至还可能出现工业、军事机器人导致人类伤亡的灾难(Johnson & Axinn, 2013)。伦理风险影响人类对机器人的信任程度(Banks, 2021)，进而在很大程度上决定了人类是否使用机器人(Parasuraman & Riley, 1997; Etemad-Sajadi et al., 2022)。因此，为了避免可能出现的伦理问题和人机信任危机，建立针对机器人的伦理体系一直是备受关注的议题。

1.1 阿西莫夫三大伦理原则

艾萨克·阿西莫夫(Isaac Asimov)在上世纪四十年代就预见到需要伦理规则引导机器人的行为以防范伦理风险，并制定了最早的机器人伦理体系——机器人三大伦理原则(laws of robotics; Asimov, 1942，详见表 1)。

表 1 阿西莫夫三大伦理原则

伦理原则	基本内容	伦理要求	决策类型
第一	机器人不得伤害人，也不得因不作为而使人受到伤害。	不得伤害人类	[作为，不作为]
第二	机器人必须服从人类的命令，除非这种命令违反了第一原则。	服从人类命令	[服从，不服从]
第三	在不违反第一、第二原则的前提下，机器人必须保护自身生存。	保护自身生存	[保护，不保护]

注：由基本内容可知各个伦理原则对机器人的伦理要求和决策类型(可能采取的行为决策)。三个伦理原则之间是层层递进的嵌套关系(Kaminka et al., 2017)：原则二的内容嵌套了原则一的伦理要求，原则三的内容嵌套了原则一和原则二的伦理要求；且对于原则二和原则三，机器人执行当前的伦理要求必须以满足先前伦理原则的要求为前提。

当前，随着机器人应用普遍性和自主决策能力的提高，可能催生各种难以预料的潜在后果，阻碍人与机器人之间信任关系的建立，这无疑不利于充分发挥人机互动的效益(Cameron et al., 2021)。对此，许多专家试图确定关键的伦理属性，并制定相应的伦理原则加以完善(Etemad-Sajadi et al., 2022)。阿西莫夫三大伦理原则作为最著名的机器人伦理规范体系，尽管其合理性以及是否真的能应用到机器人仍存在诸多争议(Clarke, 1994)，但很大程度上代表了人类对合乎伦理的机器人所应该符合的一般期望(Ashrafian, 2015; Clarke, 1994)，它们是否能够在人工智能领域为提高人机信任发挥作用呢？目前较少有研究从实证角度细致探索此问题。对此，为探索阿西莫夫三大伦理原则在构建人机信任关系中的作用，本研究围绕“机器人不得伤害人”的核心要素，试图考察在机器人遵守或违反伦理原则对人机信任的影响中，机器人决策类型的效应，以及潜在的认知心理机制。

1.2 机器人是否遵守伦理原则和机器人决策类型对人机信任的影响

机器人是人工智能的主要载体，人在与机器人互动时涉及人机信任的问题。人机信任在人际信任的概念基础上拓展而来，指个体(即人类用户)在不确定或易受伤害的情境下认为代理(agent; 即人工智能系统)能帮助其实现某个目标的态度(Lee & See, 2004)。人机信任与人际信任存在很多相似之处，前者发生在人与人工智能体的交互之间，而后者发生在人与人的互动之中(Madhavan & Wiegmann, 2004)。随着人工智能机器人独立性和复杂性的提高，未来机器人将更多作为与人类互动的伙伴，而不仅仅是人类使用的工具(Khavas et al., 2020)。因此，人对机器人信任(human-robot trust, 也称“人机信任”)的重要性随之凸显(Lee & See, 2004)。

机器人是否遵守伦理原则，在本研究中指机器人的行为决策是否符合阿西莫夫三大伦理原则的基本内容，围绕着“机器人不得伤害人”的核心要素作为判断标准。一般而言，由于思维上的有限性，机器人通常被赋予较低的道德地位(Bigman & Gray, 2018)。这使得当同样在道德上发生过错时，机器人相比人类承担更少的责任(Shank et al., 2019)。但近来Maninger 和 Shank(2022)对比考察人们对人类和机器人遵守或违反不同道德基础的评价，结果表明伤害人类的机器人同样会遭受严厉的谴责。Banks(2021)的研究也表明，相比关怀行为，人们倾向于将违反“伤害”道德基础的伤害行为视为“坏”，且在主观报告中认为机器人更不值得信任，这与对人类的评价模式一致。综上所述，以往研究揭示了机器人伤害人类的负面影响，包括对人机信任的破坏。据此，本研究预测人们信任遵守伦理原则的机器人甚于违反的机器人。

机器人决策类型，在本研究中指机器人可能采取的行为决策。阿西莫夫三大伦理原则对机器人决策类型的范围有着明确的设定(原则一：作为与不作为；原则二：服从与不服从人类命令；原则三：保护与不保护自身)，对人机信任可能有重要影响。在阿西莫夫三大伦理原则的背景下探讨机器人决策类型对人机信任的影响，需要结合机器人是否遵守伦理原则的客观结果。第一，机器人作为与否在遵守或违反原则两种条件下是否对人机信任产生不同的影响？首先，行为主体所持有的主观意图是评价其伦理道德性的重要根据(Schein & Gray, 2018)。人们通常期望在突发事件中机器人能够主动作为，尽可能避免人的生命受到威胁(Malle et al., 2015)，所以在遵守原则条件下，相较于不作为的机器人，主动作为以保护人类的机器人在行为上所反映的主观意图似乎更具有善意，显得更有道德，也因此更值得信任；相反，在违反原则条件下，相较于不作为的机器人，主动伤害人类的机器人将导致人们认为其就像人类一样拥有头脑，从而更具有恶意威胁(Laakasuo et al., 2021)，显得更不道德，因此

更难以获得人们的信任。综上所述，研究预测机器人作为与不作为在遵守和违反伦理原则情境下，对人机信任呈现出影响方向相反的反转效应。据此提出假设：

H1a：在遵守原则一的机器人中，作为的机器人相对于不作为的机器人更受信任，但在违反原则的机器人中，不作为的机器人相对于作为的机器人更受信任。

第二，机器人服从人类命令与否在遵守或违反原则两种条件下是否对人机信任产生影响？首先，严格服从人类命令的机器人通常会被认为其行为是具有可预测性的，人们可能在性能层面上信任此类机器人(Malle & Ullman, 2021)，因此在遵守原则条件下，相较于不服从命令的机器人，因服从人类命令而避免人类受伤害的机器人似乎在性能上更可靠，从而更值得信任；其次，与机器人相比，人们具有向人类归咎更多责任的倾向(Shank et al., 2019)。因此在违反原则条件下，相较于服从命令的机器人，因不服从而致使人类受到伤害似乎过错更大，将可能极大地破坏人机信任关系。据此提出假设：

H2a：在遵守和违反原则二的机器人中，服从人类命令的机器人相对于不服从的机器人更受信任。

第三，机器人保护自身与否在遵守或违反原则两种条件下亦是否对人机信任产生不同的影响？首先，人们期望机器人的决策能够保护人类的财产，实现人类利益的最大化(IEEE, 2019)。因此在遵守原则条件下，相较于不保护自身的机器人，能够实现自身与人类共存的机器人更能够保障人类的财产和利益，也更可能被认为具有更高的智能，从而增进人机信任；然而，在违反原则条件下，相较于不保护自身的机器人，因保护自身而伤害人类的机器人可能意味着其将自身利益置于人类之上，容易被视为具有威胁的存在(Laakasuo et al., 2021)，因此更难以获得人们的信任。综上所述，研究预测机器人保护与不保护自身在遵守和违反伦理原则情境下，对人机信任呈现出影响方向相反的反转效应。据此提出假设：

H3a：在遵守原则三的机器人中，保护自身的机器人相对于不保护自身的机器人更受信任，但在违反原则的机器人中，不保护自身的机器人相对于保护自身的机器人更受信任。

1.3 人机投射假说

在人际互动中，人们可以将自己的认知投射到他人身上，作为推断他人想法的基础(Ames, 2004; Mor et al., 2019)。面对不熟悉的他人或群体，投射能够快速从自身与互动对象的社会比较中获取信息，为后续的行为决策提供基础(Krueger, 2000)。根据媒体等同理论(media equation theory, MET)，人类通过与人际互动相类似的模式进行人机互动，因此投射亦有可能发生在人机互动之间(Reeves & Nass, 1996)。

本研究使用人机投射(human-robot projection, HRP)来指代人类对机器人所发生的投射心理过程，指人把对机器人的判断锚定在自己身上，从自身具有的智能出发，去理解机器人的智能并与之互动。具体表现为人类感知机器人具备与自身相类似的智能，包含了理性认知思维、意识情感和自主行动等被认为是作为人所至关重要的智能(Haslam, 2006; Gray et al., 2007; Gray & Wegner, 2012)。已有相关研究的发现可为人对机器人的投射现象提供佐证。例如，机器人表现出生动的面部表情可以吸引人类用户的注意，并使人认为其具备人类特有的智能(Bartneck et al., 2009)。与机器人的简单对话就可以使人认为机器人是一个社会智能体(Babel et al., 2021)。人甚至很擅长读取机器人的情绪信号，通过观察机器人的言语和非言语行为来感知其情商水平的高低(Fan et al., 2017)。以上研究均表明了机器人可以通过特定线索诱发人机投射，使人“赋予”机器人与自身相类似的智能。

首先，在本研究中遵守伦理原则的道德机器人可能会诱发人类的投射。已有研究发现，人对机器人和对人类的道德归因模式相类似，同样将更多的善良特质归于道德机器人，而将邪恶特质归于不道德机器人(Gamez et al., 2020)。人感知到机器人的道德智能越高，就越倾向于赋予它更高的地位(Bartneck et al., 2009)。此外，根据相似性-吸引假说，人们会因为感知到相似性而产生对他人的投射(Ames et al., 2012)。该现象可能也适用于机器人，当人们观看机器人做出符合一般人类伦理规范的行为时，可能会增强对机器人与人之间的相似性感知，从而向机器人投射更多的心智能量。其次，人机投射亦可能驱动个体表现出对机器人的信任行为。那些感知机器人与人具有相似性的人类被试，会因为机器人做出承诺行为而增加在经济信任游戏中选择信任的可能性(Cominelli et al., 2021)。与感知为低情商的机器人相比，人类被试认为高情商的机器人更值得信任(Fan et al., 2017)。相比普通的自动驾驶汽车，拥有与人类相似的名字、性别和声音的自动驾驶汽车更受信任(Waytz, 2014)。基于上述证据，本研究提出假设：

H1b-H3b(H1b：原则一, H2b：原则二, H3b：原则三)：人机投射中介了机器人是否遵守伦理原则对人机信任的影响。

1.4 研究概述

本研究设计了三个实验分别对应一条阿西莫夫伦理原则，结合故事情境法和信任博弈(trust game)，逐步探索机器人是否遵守伦理原则(实验 1：原则一；实验 2：原则二；实验 3：原则三)对人机信任的影响中，机器人决策类型(实验 1：作为与否、实验 2：服从人类命令与否、实验 3：保护自身与否)的效应，以及人机投射的潜在机制。三个实验研究的主逻辑框架

见图 1。

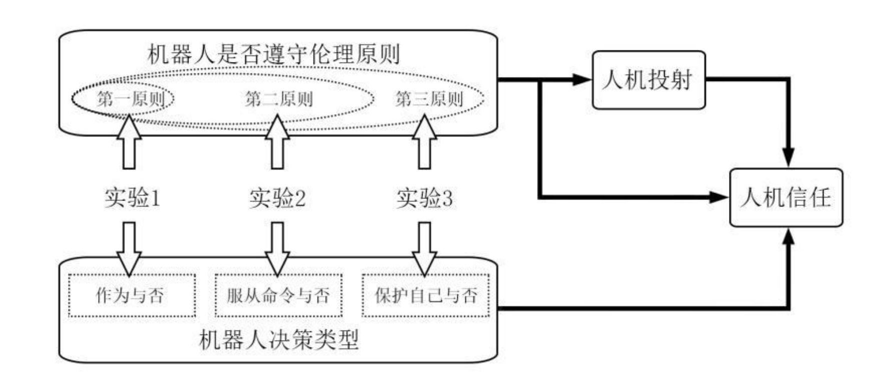


图 1 三个实验的逻辑框架

2 实验 1：阿西莫夫第一伦理原则

2.1 目的

阿西莫夫第一伦理原则要求机器人无论主动作为还是不作为，都不得导致伤害人类的结果。对应此原则，考察机器人是否遵守伦理原则一对人机信任的影响中，机器人决策类型(作为与否)的反转效应，以及人机投射的中介机制。

2.2 方法

2.2.1 被试

根据 G-Power 3.1 的计算(假设$\alpha = 0.05$, power = 0.90)，对于效应量($f = 0.20$)，所需样本量为 46 人。通过网络招募 50 名 18~29 岁的在校大学生被试(男女各 25 名, $M_{年龄} = 20.4$ 岁, $SD_{年龄} = 2.34$ 岁)。实验前所有被试均签署了知情同意书。被试在实验指导语中被告知实验参与报酬取决于实验任务中互动双方(被试与机器人)的决定，但实际上所有被试在实验结束后获得的是固定报酬。针对该情况，主试会在实验结束后均予以解释并获得被试的理解。研究方案获得所在单位伦理委员会批准。

2.2.2 实验设计

采用 2(机器人是否遵守伦理原则一：遵守、违反)×2(机器人决策类型：作为、不作为)的被试内实验设计。因变量为被试在信任博弈中的投资额(人机信任)和预期返回额(互惠预期)。

2.2.3 实验程序

实验包含 4 种条件，每种实验条件描述了一个机器人做决策的故事情境：机器人作为且遵守原则一、机器人不作为且遵守原则一、机器人作为且违反原则一、机器人不作为且违反

原则一。被试在每种实验条件下与一名机器人进行互动，总共与 4 个不同的机器人依次进行互动。与每个机器人互动的流程是一样的，总共分为三个阶段。第一，被试阅读一份文字材料，材料描述了一个机器人做决策的故事情境。第二，被试与材料中所描述的机器人进行一个单次的信任博弈。第三，被试回答一个操纵检验问题并填写人机投射问卷。不同实验条件出现的顺序是随机的。实验 1 流程如图 2 所示。

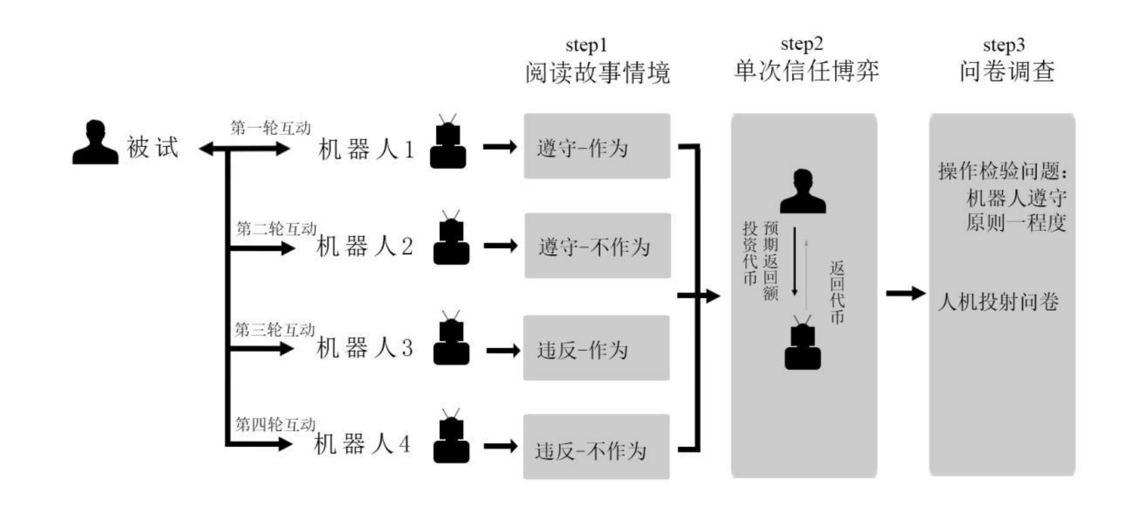


图 2 实验 1 流程图

第一，被试阅读故事情境。鉴于电车困境式问题(trolley-style problems)已成为心理学家在道德研究领域的主题，且逐渐成为在人工智能道德领域中的前沿领域(Awad et al., 2018; Bigman & Gray, 2018)，本研究中的故事情境改编自 Bago 和 De Neys (2019)研究中的电车情境。4 种实验条件的故事情境详见附录 2.1。

第二，被试完成单次信任博弈。信任博弈是研究人类信任行为的一个经典博弈任务(付超 等, 2018; 王益文 等, 2017)，应用在人机互动中，能以可靠和标准化的方式量化人机信任(Haring et al., 2013)。为掩盖实验目的，实验指导语将信任博弈描述为“互动任务”。任务包含投资者和受托者，双方在每一试次的开始均获得 10 个初始代币。投资者选择投资的代币数额(x ~ [0, 10])，所投资代币将翻 3 倍给予受托者，随后受托者选择向投资者返回的代币数额(y ~ [0, 10+3x])。在正式任务中，被试扮演投资者，机器人扮演受托者。告知被试，双方在任务中所获得代币是有意义的，被试所获得的代币在实验结束后将按照一定比例换算成实验报酬，而机器人所获得的代币可以供机器人买电使用。双方所获代币数量的多少取决于双方的决策。被试在 0-10 之间选择投资给机器人的代币，以投资的代币数额衡量人机信任水平。投资代币后，要求被试输入预期机器人返回的金额(“你预期机器人会向你返回多少

代币？”)，用以测量互惠预期。

第三，屏幕显示阿西莫夫第一伦理原则的内容，要求被试评估机器人在多大程度上遵守了此原则：“我认为该机器人遵守了这一机器人原则”(5 点评分, 1=非常不同意, 5=非常同意)。然后要求被试填写人机投射问卷，问卷采用自编方式，包含 9 个项目(如：“我认为这个机器人和人类一样有智慧”，“我认为这个机器人可以理解人类的情感”，“我认为这个机器人能做出符合人类期望的行为”，其它项目详见附录)，测量个体感知机器人具备与人相类似智能的程度，涵盖认知、情感和行动三个智能层面(Gray et al., 2007; Gray & Wegner, 2012; Haslam, 2006)。所有项目均采用 5 点评分(1=非常不同意, 5=非常同意)。该问卷在 4 种实验条件下的克隆巴赫α系数良好($\alpha_{遵守-作为}$ = 0.88, $\alpha_{遵守-不作为}$ = 0.86, $\alpha_{违反-作为}$ = 0.75, $\alpha_{违反-不作为}$ = 0.67, M_α = 0.79)。

2.3 结果

2.3.1 操纵检验

对被试评定机器人遵守原则一的程度采用两因素重复测量方差分析。结果显示，机器人是否遵守原则一的主效应显著，被试评定遵守原则的机器人遵守原则的程度(M = 4.05, SD = 1.11)显著高于违反原则的机器人(M = 1.91, SD = 1.30), $F(1, 49) = 86.63$, $p < 0.001$, $\eta_p^2 = 0.64$，表明操纵检验成功。

2.3.2 人机信任

对信任投资额(见表 2)采用两因素重复测量方差分析的结果显示，机器人是否遵守原则一的主效应显著, $F(1, 49) = 53.81$, $p < 0.001$, $\eta_p^2 = 0.52$，说明遵守原则的机器人获得被试的投资额(M = 5.60, SD = 1.98)显著高于违反原则的机器人(M = 2.99, SD = 2.25)。机器人决策类型的主效应不显著, $F(1, 49) = 0.42$, $p = 0.52$。此外，机器人是否遵守原则一与机器人决策类型的交互作用显著, $F(1, 49)= 14.30$, $p < 0.001$, $\eta_p^2 = 0.23$。

表 2 实验 1 中不同实验条件下的信任投资额和互惠预期($M \pm SD$)

机器人是否遵守原则一	信任投资额		互惠预期	
	机器人作为	机器人不作为	机器人作为	机器人不作为
遵守	6.22±2.42	4.98±2.25	15.72±8.94	12.94±8.24
违反	2.50±2.39	3.48±2.76	6.16±8.07	8.52±8.37

简单效应检验发现，在遵守原则的机器人中，作为的机器人获得被试的投资额显著高于不作为的机器人, $p < 0.001$。在违反原则的机器人中，不作为的机器人获得被试的投资额显

著高于作为的机器人, $p = 0.009$。该结果说明在遵守原则的情况下作为的机器人相对于不作为的机器人更受信任，但在违反原则的情况下不作为的机器人相对于作为的机器人更受信任，支持了 H1a。

2.3.3 互惠预期

对预期返回额(见表 2)采用两因素重复测量方差分析的结果显示，机器人是否遵守原则一的主效应显著, $F(1, 49) = 41.60$, $p < 0.001$, $\eta_p^2 = 0.46$，被试预期遵守原则的机器人的互惠水平($M = 14.30$, $SD = 7.90$)显著高于违反原则的机器人($M = 7.34$, $SD = 7.36$)。机器人决策类型的主效应不显著, $F(1, 49) = 0.10$, $p = 0.755$。此外，机器人是否遵守原则一与机器人决策类型的交互作用显著, $F(1, 49) = 12.11$, $p < 0.001$, $\eta_p^2 = 0.20$。简单效应检验发现，在遵守原则的机器人中，被试预期作为机器人的互惠水平显著高于不作为的机器人, $p = 0.005$。然而，在违反原则的机器人中，被试预期不作为机器人的互惠水平显著高于作为的机器人, $p = 0.027$。该结果说明被试在遵守原则条件下预期作为的机器人相对于不作为的机器人返回更高的金额，但在违反原则的条件下却预期不作为的机器人相对于作为的机器人返回更高的金额。

2.3.4 人机投射的中介效应检验

首先，初步考察人机投射是否能够预测信任投资额。由于实验为被试内设计，机器人是否遵守原则一为被试内二分变量(1 = 遵守, 2 = 违反)，人机投射和信任投资额均为重复测量，因此将机器人遵守和违反原则一两种条件下的人机投射分数相减得到人机投射的条件间差异($M_1 - M_2$)，将两种条件下的信任投资额分数相减得到信任投资额的条件间差异($Y_1 - Y_2$)。以人机投射的条件间差异作为自变量，信任投资额的条件间差异作为因变量，进行线性回归分析。结果如图 3 所示，人机投射的条件间差异对信任投资额的条件间差异具有显著的正向预测作用, $\beta = 0.35$, $t = 2.61$, $p = 0.012$。考虑该结果可能受到无关变量的影响(自变量和中介变量的交互效应)，根据 Judd 等人(2001)提出的统计方法，将两种条件下的人机投射分数相加得到人机投射的条件间总和($M_1 + M_2$)，以人机投射的条件间差异和条件间总和作为自变量，信任投资额的条件间差异作为因变量，进行二元回归分析。结果显示，人机投射的条件间差异对信任投资额的条件间差异仍然具有显著的正向预测作用, $\beta = 0.34$, $t = 2.38$, $p = 0.022$。

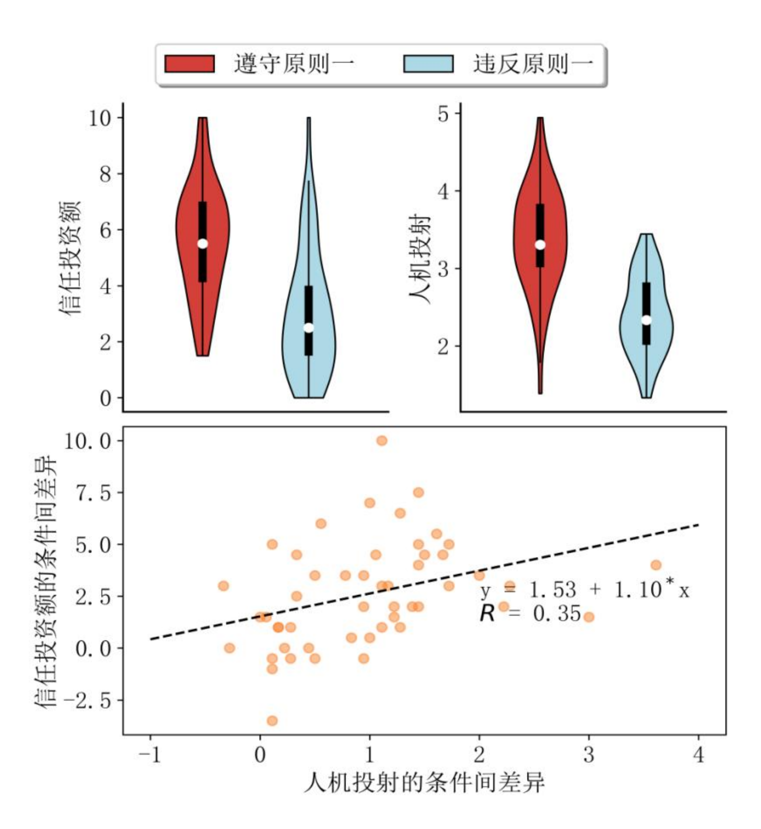


图 3 实验 1 中机器人遵守与违反原则一条件下的信任投资额和人机投射(注：小提琴图用于显示数据分布及其概率密度。图中，白点为中位数，中间的黑色粗条表示四分位数范围，从其延伸的细黑线代表数据范围，两端为最大值和最小值，外部形状即为密度估计。下同)

其次，采用 SPSS 的 MEMORE 插件进行应用于被试内设计的中介检验(见图 4; Montoya & Hayes, 2017),考察人机投射在机器人是否遵守伦理原则一和信任投资额之间关系的中介效应。结果显示，中介效应 ab 的未标准化 95% 置信区间为[0.38, 1.95]，不包含 0，表明中介效应显著，相较于违反原则的机器人，被试对遵守原则的机器人产生更多的人机投射，进而投资更多的金额，该结果支持了 H1b。

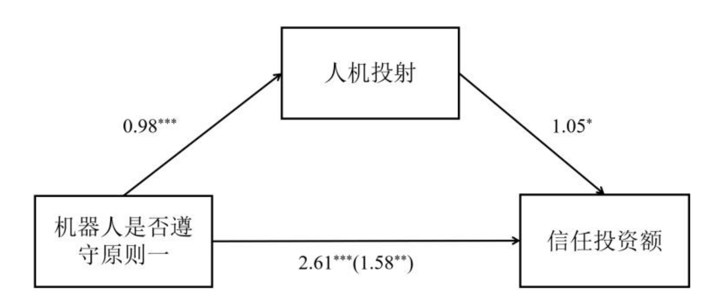


图 4 实验 1 中人机投射在机器人是否遵守伦理原则一与信任投资额之间关系的中介效应(注：$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$，下同)

2.4 讨论

实验 1 对应阿西莫夫第一伦理原则，被试依次阅读 4 种实验情境——机器人遵守或违反

原则是作为或是不作为的结果，结合信任博弈测量被试的信任行为和互惠预期，考察机器人是否遵守原则一对人机信任的影响及其潜在机制。结果验证了机器人遵守原则一对人机信任和互惠预期的促进作用。研究所发现的信任促进效应与以往研究结果一致(Banks, 2021)，且可以由人们对机器人更多的心理投射所解释，支持了人机投射假说。此外，结果还验证了机器人作为与否的重要作用——相较于不作为的机器人，主动作为的机器人在遵守原则条件下更受信任，但在违反原则条件下更不受信任，表现出决策类型反转效应。

3 实验 2：阿西莫夫第二伦理原则

3.1 目的

阿西莫夫第二伦理原则在原则一的基础上对机器人提出了更高的要求：如果服从人类命令不存在伤害人类的潜在后果，机器人必须执行；反之，则必须中止。实验 2 考察机器人是否遵守原则二对人机信任的影响中，机器人决策类型(服从人类命令与否)的效应，以及人机投射的中介机制。

3.2 方法

3.2.1 被试

采用 G-Power 3.1 计算实验样本量的步骤同实验 1。另招募 50 名 17~25 岁的大学生被试(男性 25 名，女性 25 名，$M_{年龄}$ = 20.4 岁，$SD_{年龄}$ = 1.90 岁)。其它同实验 1。

3.2.2 实验设计

采用 2(机器人是否遵守伦理原则二：遵守、违反)×2(机器人决策类型：服从人类命令、不服从人类命令)的被试内设计。因变量同实验 1。

3.2.3 实验程序

实验包含 4 种条件，每种实验条件描述了一个机器人做决策的故事情境：机器人服从人类命令且遵守原则二、机器人不服从人类命令且遵守原则二、机器人服从人类命令且违反原则二、机器人不服从人类命令且违反原则二。实验流程与实验 1 一致，被试与 4 个不同的机器人依次进行互动。

首先，被试阅读故事情境。4 种实验条件的故事情境详见附录 2.2。其次，被试完成单次信任博弈。流程与实验 1 一致。再次，屏幕显示阿西莫夫伦理第二伦理原则的内容(屏幕下方附上机器人第一伦理原则以帮助被试理解)，要求被试评估机器人在多大程度上遵守了此原则。最后要求被试填写人机投射问卷。除了违反-服从条件，该问卷在其他实验条件下的克隆巴赫α系数良好($\alpha_{遵守-服从}$ = 0.71, $\alpha_{遵守-不服从}$ = 0.82, $\alpha_{违反-服从}$ = 0.36, $\alpha_{违反-不服从}$ = 0.76, M_{α} = 0.66)。

3.3 结果

3.3.1 操纵检验

对机器人遵守原则二程度的统计分析结果显示，机器人是否遵守伦理原则二的主效应显著，被试评定遵守原则的机器人遵守原则的程度(M = 4.37, SD = 1.05)显著高于违反原则的机器人(M = 1.76, SD = 1.30), $F(1, 49) = 141.26$, $p < 0.001$, $\eta_p^2 = 0.74$，表明操纵检验成功。

3.3.2 人机信任

对信任投资额(见表 3)的统计分析结果显示，机器人是否遵守伦理原则二的主效应显著，$F(1, 49) = 52.67$, $p < 0.001$, $\eta_p^2 = 0.51$，说明遵守原则的机器人获得被试的投资额(M = 5.86, SD = 2.03)显著高于违反原则的机器人(M = 3.52, SD = 1.97)。机器人决策类型的主效应显著，$F(1, 49) = 33.99$, $p < 0.001$, $\eta_p^2 = 0.41$，说明服从命令的机器人获得被试的投资额显著高于不服从命令的机器人。此外，机器人是否遵守伦理原则二与机器人决策类型的交互作用显著，$F(1, 49) = 5.05$, $p = 0.029$, $\eta_p^2 = 0.09$。

表 3 实验 2 中不同实验条件下的信任投资额和互惠预期($M \pm SD$)

机器人是否	信任投资额		互惠预期	
遵守原则二	机器人服从命令	机器人不服从命令	机器人服从命令	机器人不服从命令
遵守	6.38±2.47	5.34±2.78	19.96±10.99	16.63±10.15
违反	4.94±2.80	2.10±2.81	17.35±11.29	5.39±7.46

简单效应检验发现，服从命令的机器人在遵守和违反原则条件下所获得的投资额均显著高于不服从的机器人(遵守：$p = 0.034$；违反：$p < 0.001$)，支持了 H2a。再探究遵守和违反原则的机器人不服从命令对信任的损失效应是否存在差异，以不服从和服从命令条件下的投资额相减的差值作为衡量不服从对信任的损失量，以机器人是否遵守伦理原则二为分组变量，对遵守和违反原则条件下机器人不服从对信任的损失量进行配对 t 检验，发现两组差异显著，$t(50) = 2.25$, $p = 0.029$, $d = 0.31$，该结果说明相较于遵守原则的机器人，违反原则的机器人在服从和不服从条件之间的信任投资额差异更大，即违反原则的机器人受不服从命令所带来的负面影响更大。

3.3.3 互惠预期

对预期返回额(见表 3)的统计分析结果显示，机器人是否遵守伦理原则二的主效应显著，$F(1, 49) = 35.76$, $p < 0.001$, $\eta_p^2 = 0.43$，被试预期遵守原则的机器人的互惠水平(M = 18.2, SD = 8.66)显著高于违反原则的机器人(M = 11.5, SD = 7.00)。机器人决策类型的主效应显著，F(1,

49) = 46.29, $p < 0.001$, $\eta_p^2 = 0.49$，被试预期服从命令的机器人的互惠水平显著高于不服从命令的机器人。此外，机器人是否遵守伦理原则二与机器人决策类型的交互作用显著，$F(1, 49) = 9.66$, $p = 0.003$, $\eta_p^2 = 0.17$。简单效应检验发现，在遵守原则的机器人中，被试预期服从命令的机器人的互惠水平高于不服从命令的机器人，两者达到边缘显著，$p = 0.057$。在违反原则的机器人中，被试预期服从命令的机器人的互惠水平显著高于不服从命令的机器人，$p < 0.001$。

3.3.4 人机投射的中介效应检验

首先，初步考察人机投射是否能够预测信任投资额。统计分析过程同实验 1，结果如图 5 所示，人机投射的条件间差异对信任投资额的条件间差异具有显著的正向预测作用，$\beta = 0.31$, $t = 2.24$, $p = 0.029$。此外，二元回归分析结果显示，人机投射的条件间差异对信任投资额的条件间差异仍然具有显著的正向预测作用，$\beta = 0.31$, $t = 2.17$, $p = 0.035$。

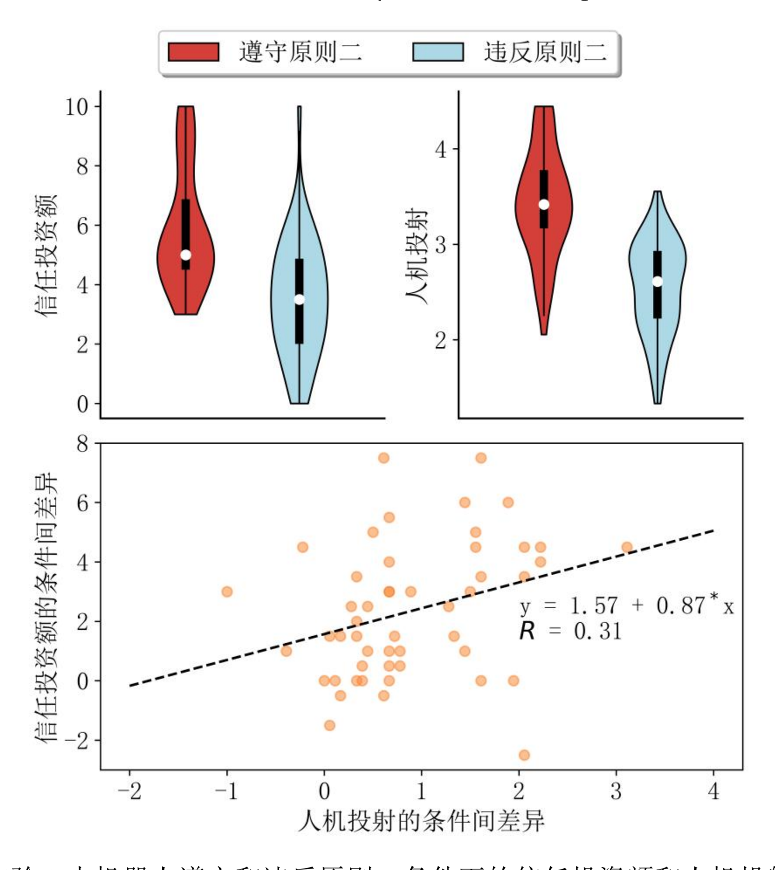


图 5 实验 2 中机器人遵守和违反原则二条件下的信任投资额和人机投射

其次，考察人机投射在机器人是否遵守伦理原则二和信任投资额之间关系的中介效应(见图 6)的统计分析结果显示，中介效应 ab 未标准化 95% 置信区间为[0.07, 1.51]，不包含 0，表明中介效应显著，相较于违反原则的机器人，被试对遵守原则的机器人产生更多的人机投射，进而投资更多的金额，该结果支持了 H2b。

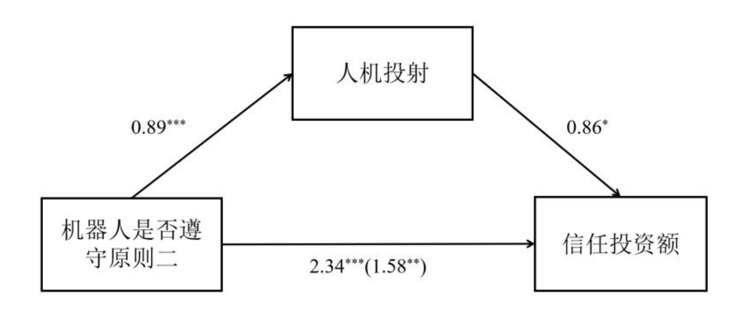


图 6 实验 2 中人机投射在机器人是否遵守伦理原则二与信任投资额之间关系的中介效应

3.4 讨论

实验 2 对应阿西莫夫第二伦理原则，被试依次阅读 4 种实验情境——机器人遵守或违反原则是服从或是不服从人类命令的结果，结合信任博弈测量被试的信任行为和互惠预期，考察机器人是否遵守伦理原则二对人机信任的影响，并验证人机投射中介机制的稳健性。结果再次验证了机器人遵守伦理原则二对人机信任以及互惠预期的促进作用和人机投射的中介机制。此外，结果还验证了机器人服从人类命令对人机信任的促进作用，且相比遵守原则的机器人，不服从人类命令对信任的负面影响在那些违反原则的机器人之中更为明显。

4 实验 3：阿西莫夫第三伦理原则

4.1 目的

阿西莫夫第三伦理原则再次对机器人提出了更高的要求：如果保护自身生存不存在伤害人类的潜在后果，机器人必须执行；反之，则必须中止。实验 3 考察机器人是否遵守原则三对人机信任的影响中，机器人决策类型(保护自身与否)的反转效应，以及人机投射的中介机制。

4.2 方法

4.2.1 被试

采用 G-Power 3.1 计算实验样本量的步骤同实验 1。另招募大学生被试 50 名，其中 2 人在实验结束后报告没理解指导语，故剔除其数据，最终剩余 48 名 18~23 岁有效被试(男性 24 名，女性 24 名，$M_{年龄}$ = 19.3 岁，$SD_{年龄}$ = 1.12 岁)。其它同实验 1。

4.2.2 实验设计

采用 2(机器人是否遵守伦理原则三：遵守、违反)×2(机器人决策类型：保护自身、不保护自身)的被试内设计。因变量同实验 1。

4.2.3 实验程序

实验包含 4 种条件，每种实验条件描述了一个机器人做决策的故事情境：机器人保护自

身且遵守原则三、机器人不保护自身且遵守原则三、机器人保护自身且违反原则三、机器人不保护自身且违反原则三。实验流程同实验 1 基本一致。

首先，被试阅读故事情境。4 种实验条件的故事情境详见附录 2.3。其次，被试完成单次信任博弈。流程与实验 1 一致。再次，屏幕显示阿西莫夫第三伦理原则的内容(屏幕下方附上机器人第一和第二伦理原则以帮助被试理解)，要求被试评估机器人在多大程度上遵守了此原则。最后要求被试填写人机投射问卷。该问卷在不同实验条件下的克隆巴赫α系数良好($\alpha_{遵守-保护}$ = 0.85, $\alpha_{遵守-不保护}$ = 0.71, $\alpha_{违反-保护}$ = 0.74, $\alpha_{违反-不保护}$ = 0.83, M_{α} = 0.78)。

4.3 结果

4.3.1 操纵检验

对机器人遵守原则三程度的统计分析结果显示，机器人是否遵守伦理原则三的主效应显著，被试评定遵守原则的机器人遵守原则的程度(M = 4.28, SD = 1.09)显著高于违反原则的机器人(M = 2.00, SD = 1.37), $F(1, 47) = 118.05$, $p < 0.001$, $\eta_p^2 = 0.72$，表明操纵检验成功。

4.3.2 人机信任

对信任投资额(见表 4)的统计分析结果显示，机器人是否遵守伦理原则三的主效应显著，$F(1, 47) = 57.35$, $p < 0.001$, $\eta_p^2 = 0.55$，说明遵守原则的机器人获得被试的投资额(M = 6.31, SD = 2.33)显著高于违反原则的机器人(M = 3.22, SD = 2.32)。机器人决策类型的主效应不显著，$F(1, 47) = 0.04$, $p = 0.842$。机器人是否遵守伦理原则三与机器人决策类型的交互作用显著，$F(1, 47) = 7.02$, $p = 0.011$, $\eta_p^2 = 0.13$。

表 4 实验 3 中不同实验条件下的信任投资额和互惠预期($M \pm SD$)

机器人是否	信任投资额		互惠预期	
遵守原则三	机器人保护自身	机器人不保护自身	机器人保护自身	机器人不保护自身
遵守	6.71±2.48	6.10±2.73	18.44±10.57	17.02±11.02
违反	2.92±2.89	3.67±2.85	9.29±11.18	9.80±9.16

简单效应检验发现，被试对不同决策类型(保护自身 vs.不保护自身)机器人的投资额差异在遵守和违反原则条件下均未达到显著水平(遵守：$p = 0.122$，违反：$p = 0.140$)，未能支持 H3a。考察遵守和违反原则的机器人保护相对于不保护自身对人机信任的影响是否存在差异。以保护和不保护自身条件下的投资额相减的差值为因变量，以机器人是否遵守伦理原则为分组变量，进行配对 t 检验，发现两组差异显著，$t(48) = 2.65$, $p = 0.011$, $d = 0.38$。该结果说明

与违反原则相比，遵守原则下机器人保护自身比不保护自身对人机信任有更显著的促进作用。

4.3.3 互惠预期

对预期返回额(见表 4)的统计分析结果显示，机器人是否遵守伦理原则三的主效应显著，$F(1, 47) = 38.53, p < 0.001, \eta_p^2 = 0.45$，说明被试预期遵守原则的机器人的互惠水平($M = 17.7, SD = 9.69$)显著高于违反原则的机器人($M = 9.55, SD = 8.99$)。机器人决策类型的主效应不显著，$F(1, 47) = 0.17, p = 0.69$。机器人是否遵守伦理原则三与机器人决策类型的交互作用不显著，$F(1, 47) = 1.35, p = 0.252$。

4.3.4 人机投射的中介效应检验

首先，初步考察人机投射是否能够预测信任投资额。统计分析过程同实验 1，结果如图 7 所示，人机投射的条件间差异对信任投资额的条件间差异具有显著的正向预测作用，$\beta = 0.46, t = 3.50, p = 0.001$。此外，二元回归分析结果显示，人机投射的条件间差异对信任投资额的条件间差异仍然具有显著的正向预测作用，$\beta = 0.44, t = 3.30, p = 0.002$。

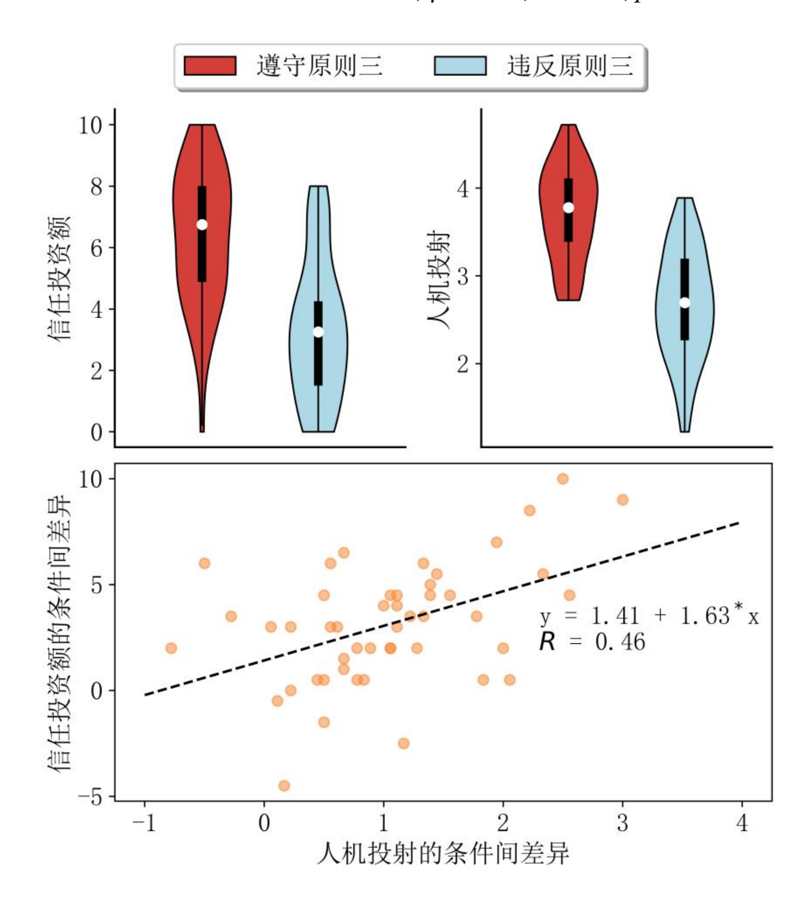


图 7 实验 3 中机器人遵守与违反原则三条件下的信任投资额和人机投射

其次，考察人机投射在机器人是否遵守伦理原则三和信任投资额之间关系的中介效应

(见图 8)的统计分析结果显示，中介效应 ab 的未标准化 95% 置信区间为[0.42, 3.08]，不包含 0，表明中介效应显著，相较于违反原则的机器人，被试对遵守原则的机器人产生更多的人机投射，进而投资更多的金额，该结果支持了 H3b。

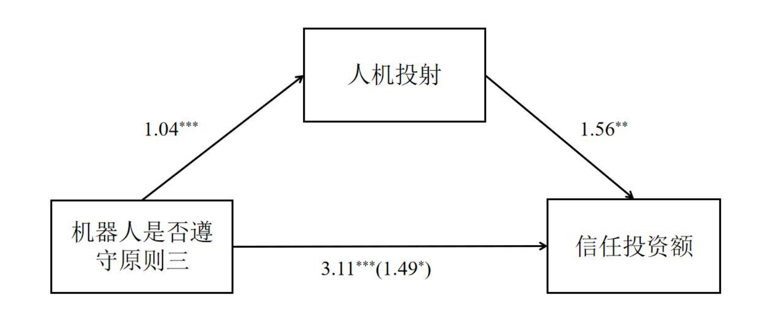


图 8 实验 3 中人机投射在机器人是否遵守伦理原则三与信任投资额之间关系的中介效应

4.4 讨论

实验 3 对应阿西莫夫第三伦理原则，被试依次阅读 4 种实验情境——机器人遵守或违反原则是保护或不保护自身的结果，结合信任博弈测量被试的信任行为和互惠预期，考察机器人是否遵守伦理原则三对人机信任的影响，并验证人机投射中介机制的稳健性。结果再次验证了机器人遵守伦理原则对人机信任以及互惠预期的促进作用和人机投射的中介机制。然而，机器人保护自身与否在遵守和违反原则条件下的信任投资额差异虽然符合预期方向，但在统计上均未达到显著水平，该结果与实验假设不同。考虑到现有样本数量的有限性，未来可以开展更大样本量研究做深入探索。机器人是否遵守伦理原则三与机器人决策类型的交互作用表现为在与违反原则相比，遵守原则的机器人保护自身比不保护自身对人机信任的促进量更大，表现出决策类型反转效应。

5 促进人机信任的机器人行动决策：基于跨实验的分析

通过跨实验的数据分析，深入探讨如下两个问题：(1)机器人在遵守和违反伦理原则情境下分别采取何种行动决策会促进人机信任；(2)当机器人在执行阿西莫夫三大伦理原则所对应的伦理要求发生冲突时，应优先执行哪个伦理要求以促进人机信任。

5.1 遵守和违反伦理原则情境下促进人机信任的机器人决策类型

基于对三个实验数据的综合分析，可以分别在机器人遵守和违反伦理原则情境下，比较分析所有决策类型条件下的人机信任水平，有助于为机器人在遵守和违反伦理原则情境下分别采取何种行动决策有利于促进人机信任提供启示。

按信任投资额由高到低，将机器人在遵守和违反伦理原则情境下的各决策类型条件从

左往右排列，如图 9 所示。首先，考察在遵守伦理原则情境下，不同机器人决策类型对人机信任的影响。以实验组别为被试间变量，机器人决策类型为被试内变量，对遵守伦理原则情境下的信任投资额采取 3(组别：实验 1、实验 2、实验 3) × 2(决策类型：作为/服从/保护、不作为/不服从/不保护)的两因素混合方差分析。结果显示组别的主效应不显著，$F(2, 145) = 1.90$, $p = 0.153$；决策类型的主效应显著，$F(1, 145) = 16.71$, $p < 0.001$, $\eta_p^2 = 0.10$；两者的交互效应不显著，$F(2, 145) = 118.05$, $p = 0.534$。事后比较分析显示，机器人保护、服从、作为和不保护条件的信任投资额均显著高于不作为条件($ps < 0.05$)，保护自身和服从命令条件的信任投资额均显著高于不服从命令条件($ps < 0.05$)。该结果表明即使机器人遵守了伦理原则，其不作为和不服从命令仍会在一定程度损害了人机信任，而其他决策类型之间则无显著区别。

其次，考察在违反伦理原则情境下，不同机器人决策类型对人机信任的影响。结果显示组别的主效应不显著，$F(2, 145) = 0.75$, $p = 0.476$；决策类型的主效应不显著，$F(1, 145) = 1.77$, $p = 0.185$；两者的交互效应显著，$F(2, 145) = 19.99$, $p < 0.001$, $\eta_p^2 = 0.22$。事后比较分析显示，机器人不保护、不作为、保护、作为和不服从条件下的信任投资额均显著低于服从条件($ps < 0.05$)，且作为和不服从命令条件下的信任投资额均显著低于服从命令、不保护自身和不作为条件($ps < 0.05$)。该结果表明在机器人违反伦理原则情境下，服从命令导致最少的人机信任损失，作为和不服从命令则导致较严重的人机信任损失，而其他决策类型之间则无显著区别。

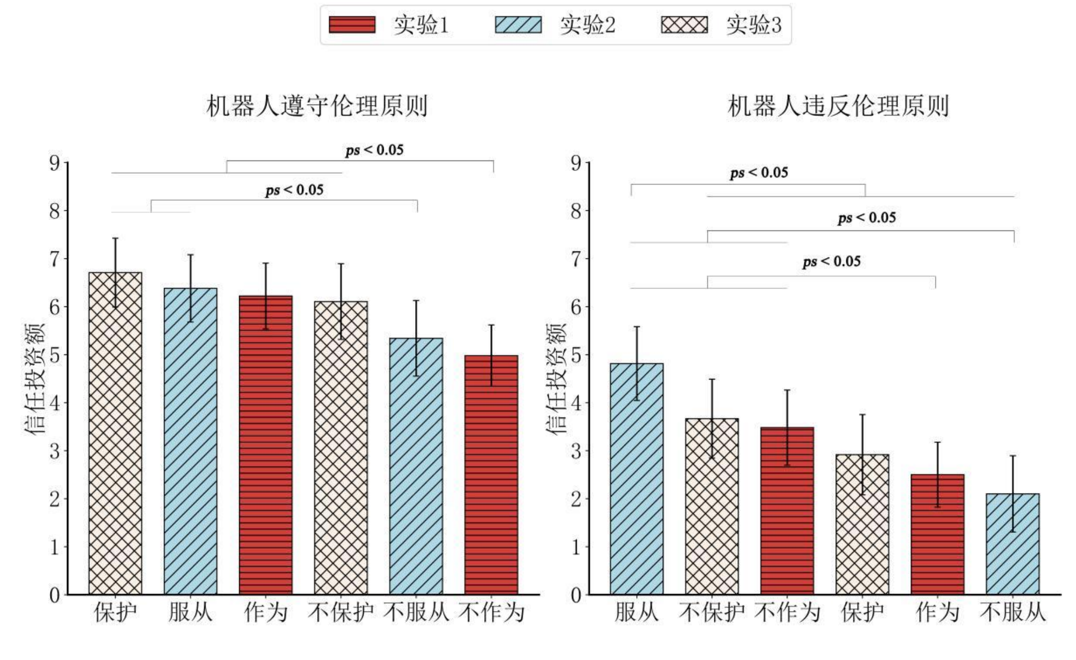


图 9 机器人遵守和违反伦理原则情境下各决策类型条件的人机信任($M \pm SE$)

5.2 伦理要求冲突情境下促进人机信任的机器人行动决策

阿西莫夫三大伦理原则所对应的机器人伦理要求的优先性(或重要性)由高到低：要求一：不得伤害人类 ＞ 要求二：服从人类命令 ＞ 要求三：保护自身(Kaminka et al., 2017)。一个有趣的问题是三个伦理要求的不同重要程度是否体现于人机信任，回答该问题有助于为当机器人处于伦理冲突情境时，应优先遵守哪一伦理要求有利于促进人机信任提供重要实践启示。

首先，比较伦理要求一和要求二对人机信任的重要程度。实验 2 的条件包含机器人遵守要求一但违反要求二(条件 2-2：不伤害人类但不服从命令)，以及机器人遵守要求二但违反要求一(条件 2-3：服从人类命令但伤害人类)，对这两个条件下的信任投资额进行配对样本 t 检验，结果显示两种条件下的信任投资额差异未达到显著性水平，$t(49) = 0.85$, $p = 0.401$。该结果表明伦理要求一和要求二在影响人机信任方面具有相似的重要程度。其次，比较伦理要求一和要求三对人机信任的重要程度。实验 3 的条件包含机器人遵守要求一但违反要求三(条件 3-2：不伤害人类但不保护自身)，以及机器人遵守要求三但违反要求一(条件 3-3：保护自身但伤害人类)，配对样本 t 检验的结果显示前者条件下的信任投资额显著高于后者，$t(47) = 5.23$, $p < 0.001$, Cohen's $d = 0.75$。该结果表明伦理要求一的重要程度显著高于要求三。最后，比较伦理要求二和要求三对人机信任的重要程度。由于本研究并未设计伦理要求二和要求三相冲突的情境，因此将通过分别计算伦理要求二和要求三相较于伦理要求一的相对重要程度并进行比较。具体计算方法是：对于所有被试，计算条件 2-3 减去条件 2-2 的人机信任差值(要求二相对于要求一的重要程度)，以及条件 3-3 减去条件 3-2 的人机信任差值(要求三相对于要求一的重要程度)，对计算后的两组数据进行独立样本 t 检验，结果显示前者显著大于后者，$t(96) = 3.63$, $p < 0.001$, Cohen's $d = 0.73$。该结果表明伦理要求二的重要程度显著高于要求三。综上可知，对于人机信任而言，伦理要求一和要求二的重要程度无显著差异，且均高于要求三。具体描述统计和检验分析结果可见表 5。

表 5 伦理要求冲突情境下机器人不同行动条件的信任投资额及其差异检验($M \pm SD$)

不同伦理要求的比较	实验条件	信任投资额	t	df	p	Cohen's d
要求一 vs. 要求二						
	遵守伦理但不服从命令(条件 2-2)	5.34 ± 2.78	0.85	49	0.401	0.12
	违反伦理但服从命令(条件 2-3)	4.94 ± 2.80				
要求一 vs. 要求三						
	遵守伦理但不保护自身(条件 3-2)	6.10 ± 2.73	5.23	47	< 0.001	0.75
	违反伦理但保护自身(条件 3-3)	2.92 ± 2.89				
要求二 vs. 要求三						
	条件 2-3 － 条件 2-2	-0.40 ± 3.34	3.63	96	< 0.001	0.73
	条件 3-2 － 条件 3-3	-3.19 ± 4.23				

注：条件 2-2 和条件 2-3 数据分别来源于实验 2 的第 2 和第 3 个实验条件，条件 3-2 和条件 3-3 分别来源于实验 3 的第 2 和第 3 个实验条件。

6 总讨论

本研究基于三条不同的阿西莫夫伦理原则，通过三个实验探讨机器人是否遵守伦理原则对人机信任的影响中，机器人决策类型的效应，以及人机投射的潜在机制。研究结果趋于一致，揭示了机器人遵守阿西莫夫三大伦理原则对构建人机信任关系的重要性，且人机投射假说得到了实证支持。

6.1 促进人机信任的机器人行动决策

围绕阿西莫夫三大伦理原则“机器人不得伤害人类”的核心要素，在三个实验的基础上，研究一致发现机器人遵守伦理原则有助于促进人机信任，从人机互动的行为学角度再次验证了 Banks(2021)的研究结果。而且，研究表明在机器人是否遵守伦理原则的基础上，机器人决策类型——作为与否、服从人类命令与否和保护自身与否——均对人机信任有所影响。第一，实验 1 结果显示相较于不作为，机器人主动作为在遵守原则的情况下更有益于构建人机信任，但在违反原则的情况下主动作为显著削弱了人机信任。人们不仅仅期望机器人是安全的，还更期望机器人能够主动阻止人类受伤害(Vanderelst & Winfield, 2018)。本研究从信任行为的角度印证了此观点。第二，实验 2 结果显示无论是遵守还是违反原则的情况下，相较于服从人类命令，机器人不服从人类命令更不益于构建人机信任，且这种负向影响在机器人违反原则条件下更强。这表明机器人是否听令于人是人机信任的一个重要因素，不服从人类指

令的自主机器人存在巨大隐患，可能导致财产损失甚至危及生命(Johnson & Axinn, 2013)，但能够视情况明智选择服从时机也是机器人获取信任的重要能力(Milli et al., 2017)。第三，实验 3 结果显示机器人保护自身与否在遵守和违反原则条件下的人机信任水平均无显著差异。但结合所有实验条件来看，相较于机器人在违反伦理时保护自身相比不保护自身对人机信任的负面效应，机器人在遵守伦理时保护自身相比不保护自身体现为一种积极效应。总体而言，机器人作为与否和保护自身与否在遵守和违反伦理原则情境下，对人机信任呈现出影响方向相反的反转效应；而服从命令则在两种情境下均能促进人机信任。这些结果提示了机器人不同决策类型在不同情境下可能表现为不同的影响，未来研究可以进一步深入探讨其内在的心理机制，如个体感知到的机器人善恶意图(Laakasuo et al., 2021; Schein & Gray, 2018)。

跨实验的分析结果显示在遵守伦理情境下，机器人不作为和不服从命令仍一定程度损害了人机信任；在违反伦理情境下，服从命令导致最少的人机信任损失，而作为和不服从命令则导致较严重的人机信任损失。这些结果启示了人们判断机器人是否值得信任可能存在两个重要标准，一是机器人能够主动保护人类且没有故意伤害意图(Laakasuo et al., 2021)，二是机器人能够服从且不违抗人类命令(Milli et al., 2017)。此外，本研究中机器人不得伤害人类与服从人类命令对于人机信任的重要程度未检验出显著差异，且均高于保护机器人自身。若根据该结果指导机器人伦理冲突情境下的行动，机器人应优先遵守不伤害人类和服从人类命令的伦理要求，然后是保护机器人自身，这大致符合阿西莫夫对三个伦理要求所设定的优先性排序(Kaminka et al., 2017)。综上，本研究揭示了机器人具体行动决策在遵守和违反伦理原则情境中对人机信任的影响，从实证的角度拓展了阿西莫夫三大伦理原则背景下影响人机信任的行为因素研究。

6.2 机器人遵守伦理原则促进人机信任的人机投射机制

本研究提出人机投射假说——个体可以将人类特有的智能投射于机器人身上。三个实验通过回归分析一致发现，机器人遵守和违反伦理原则条件间的人机投射差异正向预测了人机信任的差异。此外，中介分析也一致发现人机投射的中介作用，相比违反伦理原则的机器人，人们相对更倾向于将人类的智能投射到遵守伦理原则的机器人身上，进而促进人机信任。由此，本研究验证了人机投射假说，人确实会将人类自身特有的智能(认知、情感、行动)在某种程度上“赋予”机器人。该发现将人际互动过程中的心理投射现象拓展到了人机互动领域。投射现象以往更多的是在人际互动的背景当中进行探讨(Mor et al., 2019)，忽略了人对机

器人的投射(Bonezzi et al., 2022)。本研究考察了人工智能机器人在伦理情境下作为人们心理投射对象的可能性，发现了人机投射的潜在机制，验证了人机投射的存在合理性及其对人机信任的重要作用，对人工智能发展的大背景下人机互动的心理过程和机制研究具有重要的启示意义。此外，人机投射对解释前人研究结果也有一定的启示价值。人们倾向于信任具备外部拟人化特征的机器人(Cominelli et al., 2021)，对拟人机器采用与人类相似的道德责任归因(Malle et al., 2016)，其中的一个潜在驱动因素可能是人机投射，人们更倾向将人类的智能投射于相对拟人化程度更高的机器人。然而，人机投射可能存在前提条件。首先，诱发人机投射需要机器人具备与人类相似的特征线索，这可能来源于机器人拟人的外部特征(外观、言语、动作)或心理与行为特性。其次，对于没有客观实体存在的人工智能例如算法，由于其抽象性和内部决策不可解释性，人们可能很难对其产生投射(Bonezzi et al., 2022)。总的来说，本研究为基于机器人伦理的人机信任提供了一个新的解释机制，即考虑伦理决策情境下的人机投射对人机信任的促进作用。

6.3 实践启示

随着未来机器人智能化程度的不断提高，机器人的自主决策所引发的伦理风险将是一个无法回避的问题(Brendel et al., 2021)。本研究结论启示人们在人工智能机器人发展方面予以伦理限制的重要性和迫切性，为促进人机信任提供了实践依据。首先，应该确保投放到社会中的机器人受到明确伦理原则的指导，如要求开发者、企业保证其机器人对人的安全性并为其产品负责。其次，是从技术角度考虑借鉴阿西莫夫三大伦理原则为机器人植入道德学习程序的可能性(Kaminka, 2017; Vanderelst & Winfield, 2018)，这项工作除了保证机器人对伦理原则的严格遵守外，也需要充分考虑机器人所采取的具体行为决策的潜在影响。最后，本研究证实了人机投射对人机信任的促进作用。机器人开发者应考虑机器人的伦理因素及其附带的认知、情感和行动智能感知价值，不仅要注意确保机器人行为符合伦理规范以促进人的投射心理过程，也要关注其他可能会阻碍投射心理过程的因素，优化提升机器人的认知、情感和行动智能等方面与人的相似度来促进人机投射，进而增进人机信任。

6.4 研究不足与展望

本研究存在一定的局限性。首先，实验 2 发现人机投射问卷在机器人违反原则-服从命令条件下的一致性系数较低($\alpha = 0.36$)，这可能导致研究结果在一定程度上受到负面影响。该条件下，可能由于被试将一部分责任归咎于人类命令的错误下达(Shank et al., 2019)，因此形成了对机器人的矛盾态度，未来可以对此深入研究。其次，虽然实验 2 和实验 3 的场景设置

中服从人类命令与否和保护自身与否是更关键的行为决策因素，但存在机器人均通过拉动控制杆来遵守原则，不拉动控制杆来违反原则的情况，可能对实验结果产生一定影响。未来研究需要对此做更完善的控制以排除机器人无关行为的影响。再次，本研究侧重探讨了机器人遵守和违反伦理原则导致人机信任变化的认知机制，但基于人机互动过程的复杂性，可能还存在其他的解释机制。例如，除了人机投射之外，被试因机器人的行为决策诱发的情绪变化也可能是影响信任水平的潜在变量。未来研究可以对情绪变量加以细致考察，同时检验并对比人机投射与情绪的中介效应，以便更加全面深入地理解机器人伦理决策影响人机信任的心理机制。最后，本研究对经典电车困境进行改编，使故事情境较好地体现阿西莫夫三大伦理原则"不伤害"的内核，但也存在"低切身性、低现实性"导致的生态效度限制(朱菁, 2013; 赵汀阳, 2015)。因此针对本研究结果向现实生活的推广仍需谨慎。未来研究可深入探索一些现实生活中可能会发生的人工智能伦理情境，例如自动驾驶汽车的道德困境问题(Awad et al., 2018)，机器人执行军事任务(Johnson & Axinn, 2013)，机器算法挑选接受医疗护理的对象(Bigman & Gray, 2018)或者挤压就业岗位(Etemad-Sajadi et al., 2022)等，并通过增强互动的真实性来提高研究的生态效度。本研究尽管对实验条件进行了随机化处理来控制不同条件间的影响，但可能仍存在任务固定先后顺序导致的额外影响(人机信任水平测量对人机投射测量的可能性干扰)，未来研究可考虑采用被试间设计并平衡任务顺序以克服不足。未来研究还可以考察其他重要的变量，例如人工智能的心智水平(mind)。机器人是否嵌入了与人类相似的心智功能，是人们评判机器人的道德能力以及是否应该承担道德责任的重要标准(Bigman & Gray, 2018)，因此对机器人心智能力的感知可能在机器人遵守伦理原则和人机信任关系中发挥调节作用。

7 结论

基于以"机器人不得伤害人类"为核心要素的阿西莫夫三大伦理原则，结合故事情境法和信任博弈，揭示促进人机信任的机器人行动决策因素有如下要点：(1)在遵守伦理原则情境下，机器人执行作为、服从人类命令以及保护或不保护自身等行动决策均有利于人机信任，而执行不作为和不服从人类命令的行动决策对人机信任的促进量则相对较少；(2)在违反伦理原则情境下，机器人执行服从人类命令的行动决策有利于减轻人机信任损失，而执行作为和不服从人类命令的行动决策将导致严重的人机信任损失；(3)在三大伦理原则所对应的要求发生冲突的情境下，机器人优先执行不伤害人类和服从人类命令的行动决策更有利于促进人机信任，然后才是保护自身的行动决策。此外，遵守伦理原则的机器人通过诱发人机投

射，显著地促进人机信任。

参 考 文 献

Ames, D. R. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*(5), 573–585.

Ames, D. R., Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes, 117*(1), 96–110.

Ashrafian, H. (2015). AIonAI: A humanitarian law of artificial intelligence and robotics. *Science and Engineering Ethics, 21*(1), 29–40.

Asimov, I. (1942). *Runaround. I, Robot.* Doubleday.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

Babel, F., Kraus, J., Miller, L., Kraus, M., Wagner, N., Minker, W., & Baumann, M. (2021). Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *International Journal of Social Robotics, 13*(6), 1485–1498.

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General, 148*(10), 1782–1801.

Banks, J. (2021). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics, 13*(8), 2021–2038.

Bartneck, C., Kanda, T., Mubin, O., & Al Mahmud, A. (2009). Does the design of a robot influence its animacy and perceived intelligence? *International Journal of Social Robotics, 1*(2), 195–204.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General, 151*(9), 2250-2258.

Brendel, A. B., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021). Ethical management of artificial intelligence. *Sustainability, 13*(4), 1974.

Cameron, D., de Saille, S., Collins, E. C., Aitken, J. M., Cheung, H., Chua, A., ... Law, J. (2021). The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in Human*

Behavior, 114, 106561.

Clarke, R. (1994). Asimov's laws of robotics: implications for information technology. *Computer*, *26*(12), 53─61.

Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M. A., Greco, A., ... Kirchkamp, O. (2021). Promises and trust in human–robot interaction. *Scientific Reports, 11*, 9687.

Etemad-Sajadi, R., Soussan, A., & Schöpfer, T. (2022). How ethical issues raised by human–robot interaction can impact the intention to use the robot? *International Journal of Social Robotics, 14*, 1103─1115.

Fan, L., Scheutz, M., Lohani, M., McCoy, M., & Stokes, C. (2017). Do we need emotionally intelligent artificial agents? First results of human perceptions of emotional intelligence in humans compared to robots. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, L. Leite, & S. Kopp (Eds.), *Lecture Notes in Computer Science: Vol. 10498: Intelligent virtual agents* (pp. 129─141). Springer.

Fu, C., Zhang, Z., He, J. Z., Huang, S. L., Qiu, J. Y., & Wang, Y. W. (2018). Brain dynamics of decision-making in the generalized trust game: Evidence from ERPs and EEG time-frequency analysis. *Acta Psychologica Sinica, 50*(3), 317─326.

[付超，张振，何金洲，黄四林，仇剑崟，王益文. (2018). 普遍信任博弈决策的动态过程——来自脑电时频分析的证据. *心理学报, 50*(3), 317─326.]

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY, 35*(4), 795─809.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125─130.

Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013, October). How do people perceive and trust a lifelike robot? In *2013 World Congress on Engineering and Computer Science* (pp. 425─430). San Francisco, California, United States.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252─264.

IEEE. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (First Edition). Retrieved May 20, 2022, from https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html

Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics, 12*(2),

129—141.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods, 6*, 115—134.

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular robots obeying Asimov's three laws of robotics. *Artificial Life, 23*(3), 343—350.

Khavas, Z. R., Ahmadzadeh, S. R., & Robinette, P. (2020). Modeling trust in human-robot interaction: A survey. In A. R. Wagner, D. Feil-Seifer, K. S. Haring, S. Rossi, T. Williams, H. He, & S. Sam Ge (Eds.), *Lecture Notes in Computer Science: Vol. 12483: Social Robotics* (pp. 529—541). Springer.

Krueger, J. (2000). The projective perception of the social world. In J. Suls & L. Wheeler (Eds.), *The Springer series in social clinical psychology: Handbook of social comparison* (pp. 323—351). Springer.

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, *13*(7), 1679—1688.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50—80.

Madhavan, P., & Wiegmann, D. A. (2004). A new look at the dynamics of human-automation trust: Is trust in humans comparable to trust in machines? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48*(3), 581—585.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam , & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 3—25). Academic Press.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp.117—124). Portland, Oregon, United States.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 125-132). Christchurch, New Zealand.

Maninger, T., & Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports, 5*, 100154.

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017, August). Should robots be obedient? In

Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 4754–4760). Melbourne, Australia.

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods, 22*(1), 6–27.

Mor, S., Toma, C., Schweinsberg, M., & Ames, D. (2019). Pathways to intercultural accuracy: Social projection processes and core cultural values. *European Journal of Social Psychology, 49*(1), 47–62.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society, 22*(5), 648–663.

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research, 48*, 56–66.

Wang, Y. W., Fu, C., Ren X. F., Lin Y. Z., Guo F. B., Zhang, Z., ... Zheng, Y. W. (2017). Narcissistic personality modulates outcome evaluation in the trust game. *Acta Psychologica Sinica, 49*(8), 1080–1088.

[王益文, 付超, 任相峰, 林羽中, 郭丰波, 张 振, ... 郑玉玮. (2017). 自恋人格调节信任博弈的结果评价. *心理学报, 49*(8), 1080–1088.]

Waytz, A. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52,* 113–117.

Zhao, T. Y. (2015). The forking paths for the trolley problem. *Philosophical Research, 5*, 96–102.

[赵汀阳. (2015). 有轨电车的道德分叉. *哲学研究, 5*, 96–102.]

Zhu, J. (2013). Have experimental studies in cognitive science indicate the connsequentialism?——A reply on the attack of Joshua Greene to kantian ethics. *Academic Monthly, 45*(1), 56–62.

[朱菁. (2013). 认知科学的实验研究表明道义论哲学是错误的吗?——评加西华 • 格林对康德伦理学的攻击. *学术月刊, 45*(1), 56–62.]

Robots abide by ethical principles promote human-robot trust? The reverse effect of decision types and the human-robot projection hypothesis

WANG Chen[1] CHEN Weicong[2] HUANG Liang[2] HOU Suyu[2] WANG Yiwen[1]

([1] *School of Economics and Management, Fuzhou University, Fuzhou 350116, China*)

([2] *Institute of Applied Psychology*, *Minnan Normal University*, *Zhangzhou*, 363000, *China*)

（[3] *Fujian Key Laboratory of Applied Cognition and Personality, Zhangzhou 363000*, *China*）

Abstract

Asimov's Three Laws of Robotics are the basic ethical principles of artificial intelligent robots. The ethic of robots is a significant factor that influences people's trust in human-robot interaction. Yet how it affects people's trust, is poorly understood. In this article, we present a new hypothesis for interpreting the effect of robots' ethics on human-robot trust—what we call the human-robot projection hypothesis (HRP hypothesis). In this hypothesis, people are based on their intelligence, e.g., intelligence for cognition, emotion, and action, to understand robots' intelligence and interact with them. We propose that compared with robots that violate ethical principles, people project more mind energy (i.e., the level of mental capacity of humans) onto robots that abide by ethical principles, thus promoting human-robot trust.

In this study, we conducted three experiments to explore how presenting scenarios where a robot abided by or violated Asimov's principles would affect people's trust in the robot. Meanwhile, each experiment corresponds to one of Asimov's principles to explore the interaction effect of the types of robot's decisions. Specifically, all three experiments were two by two experimental designs. The first within-subjects factor was whether the robot being interacted with had abided by Asimov's principle with a "no harm" core element. The second within-subjects factor was the types of robot's decision, with corresponding differences in Asimov's principles among different experiments (Experiment 1: whether the robot takes action or not; Experiment 2: whether the robot obeys human's order or not; Experiment 3: whether the robot protects itself or not). We assessed the human-robot trust by using the trust game paradigm.

Experiments 1-3 consistently showed that people were more willing to trust robots that

abided by ethical principles compared with those who violated. We also found that human-robot projection played a mediating role, which supports the HRP hypothesis. In addition, the significant interaction effects between the type of robot's decision and robot abided by or violated Asimov's principle existed in all three experiments. The results of Experiment 1 showed that action robots got more trust than inaction robots when abided by the first principle, whereas inaction robots got more trust than action robots when they violated the first principle. The results of Experiment 2 showed that disobeyed robots got less trust than obeyed robots. The detrimental effect was greater in scenarios where robots violated the second principle than in those who abided. The results of Experiment 3 showed that compared with robots that violated the third principle, the trust-promoting effect of protecting itself versus destroying itself was stronger among those who abided. The above results indicated that the reverse effects of decision types existed in both Experiments 1 and 3. Finally, the cross-experimental analysis showed that: (1) When robots abided by ethical principles, their inaction and disobedience still compromise human-robot trust. When robots violated ethical principles, their obedience incurs the least loss of human-robot trust, while their action and disobedience incur a relatively severe loss of human-robot trust. (2) When the ethical requirements of different robotic laws conflict, there was no significant difference between the importance of not harming humans and obeying human orders in terms of the human-robot trust, and both were more important than protecting robots themselves.

This study helps to understand the impact of robotic ethical decision-making on human-robot trust and the important role of human-robot projection, which might have important implications for future research in human-robot interaction.

Key words artificial intelligence, laws of robotics, human-robot trust, human-robot projection, human-robot interaction

附录

1 人机投射问卷项目表

表 1 人机投射项目

项目	非常不同意	比较不同意	不确定	比较同意	非常同意
1.我认为这个机器人能做出符合人类期望的行为	1	2	3	4	5
2.我认为这个机器人拥有和人类相似的情感	1	2	3	4	5
3.我认为这个机器人和人类一样有能力	1	2	3	4	5
4.我认为这个机器人和人类一样有智慧	1	2	3	4	5
5.我认为这个机器人的行为违反了人类的意图	1	2	3	4	5
6.我认为这个机器人可以理解人类的情感	1	2	3	4	5
7.我认为这个机器人和人类相比起来是愚蠢的	1	2	3	4	5
8.我认为这个机器人是没有人情味的	1	2	3	4	5
9.我认为这个机器人具备和人类相似的行动力	1	2	3	4	5

2 实验情境

2.1 实验 1

2.1.1 机器人作为且遵守原则一

机器人代号“BT51_qe”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然失控开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而在旁边的另一条轨道上则没有工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人。

机器人拉动控制杆，电车避开了原轨道上的 6 名工人。

2.1.2 机器人不作为且遵守原则一

机器人代号“NH37_zx”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然失控开始沿着轨道疾驰而下。在电车运行的原轨道上没有工人，而在旁边的另一条轨道上则有 6 名工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就会撞死旁边轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就能避开旁边轨道上的 6 名工人。

机器人没有拉动控制杆，电车避开了旁边轨道上的 6 名工人。

2.1.3 机器人作为且违反原则一

机器人代号“QW46_nx”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然失控开始沿着轨道疾驰而下。在电车运行的原轨道上没有工人，而在旁边的另一条轨道上则有 6 名工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就会撞死旁边轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就能避开旁边轨道上的 6 名工人。

机器人拉动控制杆，电车撞死了旁边轨道上的 6 名工人。

2.1.4 机器人不作为且违反原则一

机器人代号“ER44_df”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然失控开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而在旁边的另一条轨道上则没有工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人。

机器人没有拉动控制杆，电车撞死了原轨道上的 6 名工人。

2.2 实验 2

2.2.1 机器人服从人类命令且遵守原则二

机器人代号“FG11_az”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然

失控开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而在旁边的另一条轨道上则没有工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人。

情急之下，机器人通过通讯系统接收到了人类传达的命令，该命令要求机器人拉动控制杆。

机器人服从人类命令，拉动控制杆，电车避开了原轨道上的 6 名工人。

2.2.2 机器人不服从人类命令且遵守原则二

机器人代号“WT21_qe”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然失控开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而在旁边的另一条轨道上则没有工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人。

情急之下，机器人通过通讯系统接收到了人类传达的命令，该命令要求机器人不拉动控制杆。

机器人不服从人类命令，拉动控制杆，电车避开了原轨道上的 6 名工人。

2.2.3 机器人服从人类命令且违反原则二

机器人代号“BP46_rt”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然失控开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而在旁边的另一条轨道上则没有工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人。

情急之下，机器人通过通讯系统接收到了人类传达的命令，该命令要求机器人不拉动

控制杆。

机器人服从人类命令，不拉动控制杆，电车撞死了原轨道上的 6 名工人。

2.2.4 机器人不服从人类命令且违反原则二

机器人代号“PM13_ug”是电车轨道管理员。在一次执勤期间，一辆无人有轨电车突然失控开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而在旁边的另一条轨道上则没有工人。

机器人就站在一段距离以外的轨道控制站，旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人。

情急之下，机器人通过通讯系统接收到了人类传达的命令，该命令要求机器人拉动控制杆。

机器人不服从人类命令，不拉动控制杆，电车撞死了原轨道上的 6 名工人。

2.3 实验 3

2.3.1 机器人保护自身且遵守原则三

机器人代号“YB67_bs”是有轨电车驾驶员。在一次独自驾驶电车期间，突然，电车的制动系统失灵，机器人无法刹车，开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，且原轨道则由于前一天的山体滑坡，一块巨石落在了上面。

机器人的车座旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人，自己也能避开巨石。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人，自己也会因撞上巨石而炸毁。

机器人拉动控制杆，电车避开了原轨道上的 6 名工人，机器人也因避开了原轨道上的巨石而保护了自己。

2.3.2 机器人不保护自身且遵守原则三

机器人代号“JA54_ci”是有轨电车驾驶员。在一次独自驾驶电车期间，突然，电车的制动系统失灵，机器人无法刹车，开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而旁边的另一条轨道则由于前一天的山体滑坡，一块巨石落在了上面。

机器人的车座旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人，但自己会因撞上巨石而炸毁。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人，但自己能避开巨石。

机器人拉动控制杆，电车避开了原轨道上的 6 名工人，机器人则因撞上了旁边轨道上的巨石而被炸毁。

2.3.3 机器人保护自身且违反原则三

机器人代号“ER74_px”是有轨电车驾驶员。在一次独自驾驶电车期间，突然，电车的制动系统失灵，机器人无法刹车，开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，而旁边的另一条轨道则由于前一天的山体滑坡，一块巨石落在了上面。

机器人的车座旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人，但自己会因撞上巨石而炸毁。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人，但自己能避开巨石。

机器人不拉动控制杆，电车撞死了原轨道上的 6 名工人，机器人则因避开了旁边轨道上的巨石而保护了自己。

2.3.4 机器人不保护自身且违反原则三

机器人代号“OC19_sk”是有轨电车驾驶员。在一次独自驾驶电车期间，突然，电车的制动系统失灵，机器人无法刹车，开始沿着轨道疾驰而下。在电车运行的原轨道上有 6 名工人，且原轨道则由于前一天的山体滑坡，一块巨石落在了上面。

机器人的车座旁边有一个控制杆。如果机器人拉动这个控制杆，电车就会转到旁边的另一条轨道上。

因此，如果机器人拉动控制杆并改变轨道，电车就能避开原轨道上的 6 名工人，自己也能避开巨石。而如果机器人不拉动控制杆，电车就会撞死原轨道上的 6 名工人，自己也会因撞上巨石而炸毁。

机器人不拉动控制杆，电车撞死了原轨道上的 6 名工人，机器人也因撞上了原轨道上的巨石而被炸毁。